annotation associated with the GenBank entry. However, the same annotation continues that the entry was updated on Aug 11, 1997. The Examiner indicated that the Action is premised on an assumption that the relied upon sequence was published on Oct 6, 1995.

The creation date of an EMBL or GenBank record is not the public availability date. The creation date is the date the record was originally created. Frequently, these records are maintained in secrecy until a predetermined publication or patent filing date is effected. Furthermore, the creation date does not often reflect the record as subsequently accessed. Like most electronic databases, Genbank and EMBL are constantly updating, amending, annotating and otherwise supplementing their records. These newer "editions" retain the creation date of the original record, but were obviously not in existence at that date. Here, the Examiner seeks to rely on a creation date for a record that could not logically have existed on that creation date. A document (electronic or otherwise) that makes explicit reference to dates and events in 1997 could not logically have been "published" or made "publically available" in 1995. This rejection is akin to citing a year 2002-updated article in the Encyclopedia Brittanica and relying on the encyclopedia's year 1768 original publication date.

Highlighted copies of EMBL and GenBank database information for submitters (including information on withholding public availability of records after submission and record creation) is attached. Also attached is a Sample GenBank Record explaining that even the date of last modification may not correspond to the release date (p.6).

The Examiner apparently seeks to shift the burden to Applicants to prove that Rose et al. is not prior art. We believe this position is as untenable as imposing an inherently impossible proof on the Applicants, and misconstrues the duty of the Examiner, which is to allow our claims unless he can establish a prima facie case of non-patentability, which includes showing that the cited art is prior art. Here, the uncontroverted evidence unequivocally demonstrates that what is cited was not what was publically available as of Oct 6, 1995.

In any event, the entire yeast genome had been largely sequenced prior to the filing of our patent application, including the identification of thousands of ORFs which were not even known to encode functional mRNA. Even if these ORFs contained an identical or substantially identical sequence, the claimed compositions would be neither anticipated nor obvious. First, even if a yeast chromosome sequence is determined (and it appears that a sequence encoding Afc1p (SEQ ID NO:2) is found on the yeast X chromosome), our claims do not encompass any chromosome. Second, our claims require that the coding sequence be operatively joined to a promoter. In the absence of any evidence for function, there would be no motivation to select out one of the thousands of yeast ORFs of unknown function, isolate what may or may not be a coding

sequence, and operatively join it to a promoter.

(b) The Examiner's rejection of claims 35, 37-38 and 43, 45-46 over Lye et al. (GenBank Accession No. Z49260) in view of Nozaki et al. (US Pat No. 4,997,767) is not in compliance with the notice requirement of 35USC132, which requires reason and information and references useful in judging the propriety of the rejection.

The art rejection applied to claims 35-38 and 43-46 (reciting SEQ ID NOS:3 and 4) relies on Lye, et al. (GenBank Database, Accession No. Z49260), which is also dated Aug 11, 1997, more than a year after our Aug 7, 1996 priority date, and is hence not prior art. Instead of a publication date, the Examiner appears to rely on a purported unpublished submission date. This is improper; if a database entry does not recite a publication date, it can not be relied upon as prior art; see MPEP2128. The Action offers no evidence that the relied upon sequence was published at any time prior to Aug 11, 1997.

In any event, the entire yeast genome had been largely sequenced prior to the filing of our patent application, including the identification of thousands of ORFs which were not even known to encode functional mRNA. What Lye discloses are computer predictions of thousands of possible CDS regions. A computer is programmed to input raw genomic sequence, select all possible CDS regions over 100 codons, and then exclude those that are more than 50% overlapped by a larger predicted CDS. The authors promise that CDS regions of the initial dataset subsequently eliminated by the algorithm are nevertheless "available upon request." In addition, the disclosure provides algorithm-predicted PROSITE database matches, though the authors caution that some of these may be "fortuitous".

Lye does not disclose any gene or gene product, but the results of a first run effort to sequence the entire XIII chromosome of Saccharomyces cerevisiae. That natural yeast XIII chromosome is, of course, prior art, and Lye provides no more than an inherent property of that chromosome - its sequence. Lye discloses no more than raw genomic data weighted by a computer for thousands of possible genes and genetic elements. The Examiner uses our own disclosure to select out one of these and uses our own disclosure to provide motivation to recombine it in an expression vector. In the absence of any evidence for function, there would be no motivation to select out one of the thousands of yeast ORFs of unknown function, isolate what may or may not be a coding sequence, and operatively join it to a promoter in an expression vector, as expressly required by our claims.

Absent a prior art suggestion that SEQ ID NO:1 or 3 encodes a protein of determined function sufficient to motivate the isolation, cloning and expression of such SEQ ID NO using

techniques such as those of the cited Nozaki et al. (US Pat No 4,997,767) and Sambrook, J. et al. (Mol. Cloning, Cold Spring Harbor Press, p. 16.3-16.16) references, the claims are in compliance with 35USC102 and 103.

We hereby petition for and authorize charging to our Deposit Account No. 19-0750 all necessary extensions of time. The Commissioner is hereby authorized to charge any necessary fees associated with this communication to our Deposit Account No. 19-0750 (order no. B96-021-3).

Respectfully submitted,
SCIENCE & TECHNOLOGY LAW GROUP

Richard Aron Osman, Ph.D., Reg. No. 36,627
Tel: (650) 343-4341; Fax: (650)343-4342

encl. EMBL database information (7p)
GenBank database information (5p)
Sample GenBank Record (17p)

EMBL
Europ an Bioinformatics Institute

Search Nucleotide sequences for Go ? Site search: Go ?

EMBL Database
Access
Submission
Documentation
Group info
Contact
News
EBI Home

The EMBL Nucleotide Sequence Database

Information for Submitters

Get this document in .doc or .mcw format

Introduction

Submission of sequence information to the nucleotide sequence database prior to publication has become standard practice. A unique accession number is assigned by the database which permanently identifies the sequence submitted. The database accession number should be included in the manuscript, preferably on the first page of the journal article, or as required by individual journal procedures. This procedure ensures availability and distribution of new sequence data in a timely fashion.
Note: It is only necessary to submit to one database, without regard to where the sequence will be published. Data are exchanged between EMBL, GenBank and DDBJ on a daily basis.

[index]

Checking sequences for vector contamination

To assist submitters the EBI provides a vector screening service using the latest implementation of th BLAST algorithm and a special sequence databank known as EMVEC. EMVEC is an extraction of sequences from the SYNthetic division of EMBL containing more than 2000 sequences commonly used in cloning and s quencing

experiments. EMVEC is by no means a complete vector databank but it is representative of the kind of material used in modern sequencing. The databank will be updated with each release of EMBL and mad publicly available on the EBI's FTP (ftp.ebi.ac.uk) server.

The interactive WWW service can be found at:
http://www2.ebi.ac.uk/blastall/vectors.html

[index]

How to submit data to the EMBL Nucleotide Sequence Database

Webin - WWW Nucleotide Sequence Submissions

Webin is EBI's preferred submission medium. Webin guides the user through a sequence of WWW forms allowing interactive submission of sequence data and descriptive information to the EMBL database. All the information required to create a database entry will be collected during this process:

1. Submitter Information
2. Release Date Information
3. Sequence Data, Description and Source Information
4. Reference Citation Information
5. Feature Information (e.g. coding regions, regulatory signals etc.)

Submitters are able to modify and view their data prior to submission in the format in which it will be finally published in the EMBL database. Using Webin is the quickest way to get your accession numbers assigned.

Sequin

Sequin is the latest multi-platform (Mac/PC/Unix) stand-alone software tool developed by the NCBI for submitting entries to the EMBL, GenBank, or DDBJ sequence databases. The Sequin program, along with detailed downloading and installation instructions plus general information are available from the EBI via WWW browser, anonymous FTP and from the file server.

Data Submission Form

The Data Submission Form has been updated (December 1998) and is available here. Since June 1999 The EMBL Nucleotide Database does not accept submissions on the old Data Submission Form. If received such submissions will be returned and will have to be resubmitted either on the new Data Submission Form or via Webin.

[index]

What to submit to the EMBL Nucleotide Sequence Database

Direct submissions to the EMBL database are reviewed by nucleotide database curators, but the ultimate responsibility for the accuracy and quality of the information lays with the submitter. Please check the EMBL annotation examples we provide to ensure that you include all important biological features into your submission. To make your entries easily retrievable by other scientists working in the same field, please follow the various nomenclature standards (i.e. gene nomenclature, product nomenclature) set by corresponding organisations.
The small collection of useful nomenclature links is available here.

WWW

Data submitted via the WWW submission system will contain the required components, although we may contact the author concerning details.

Sequin

Sequin output, generated by selecting the 'Prepare Submission' menu option in computer-readable form by electronic mail Submission Form. A completed data submission form for each submitted sequence plus the continuous sequence(s) by e-mail.

[index]

How to send data to the EMBL Nucleotide Sequence Database

Data can be sent to the Nucleotide Sequence Database via:

Webin. Data submitted via the WWW submission system will be automatically transmitted to the EBI database staff on successful completion of the interactive forms.

Electronic mail to DATASUBS@EBI.AC.UK . Note: The EMBL nucleotide database no longer accepts Authorin submissions. Submitters who previously used Authorin should now use one of three submission mechanisms - Webin, Sequin or Data Submission Form.

[index]

How long will it take to get an accession number?

We will process data submissions within 2 working days of receipt (5 working days for bulk submissions) and send authors notification of either what accession number(s) their data have been assigned or what additional information is needed.

To minimise the time it takes to get an accession number:

1. Use Webin - this is the quickest way to get your accession numbers assigned.
2. Be sure that submissions include all the necessary information.
3. Check the data for inconsistencies/errors (e.g., a stop codon in the middle of a coding region, sequence length same as stated on form).
4. Be sure to include either a computer network address or a telefax number. If this information is not provided, notification of accession numbers will be sent by regular post

[index]

Data confidentiality and release dates

Authors will be asked whether their submitted data can be made available to the public immediately or whether they should be withheld until an author-specified date. Data are never withheld after publication.

[index]

Bulk Submissions

For researchers wishing to submit 25 or more related sequences (e.g. the sam gene sequenced in a large number of different organisms), WEBIN offers a bulk submission procedur . This alternative path through WEBIN allows submitters to create one representative sequence entry. By instructing EMBL curators which of the entries' features differ between each sequence, minimal template WEBIN forms are customised to fit the exact requirements of that particular set of sequences. The bulk procedure is highly efficient and less time consuming for the submitter, who no longer has to duplicate information. The procedure also ensures that EMBL curators process related data together and consistently. Because there are fewer forms, just one form per 10 sequences, bulk submissions are also much faster over slow networks. Prospective submitters should note that an EMBL curator will review the initial representative sequence within five working days. Submitters are then notified by e-mail that the templates are ready and may then complete their submission.

Please contact database staff if you require further information.
e-mail: datasubs@ebi.ac.uk

[index]

Genome Project Submissions

For groups producing large volumes of nucleotide sequence data over an expended period, submission accounts can be established with the EBI.
Such groups include the genome sequencing and mapping projects.
A submission protocol is agreed upon and database entries produced at the research site can be deposited and updated directly by the originating group via FTP or electronic mail. Each submission account is 'curated' by EBI biologists, who check to ensure that new entries follow database annotation conventions and are consistent with other entries from the same project. The curator also serves as an informed liaison between the sequencing group and the database.

Genome Project Submission Account guidelines are available here.

We welcome enquiries from any researcher who thinks they may be a suitable candidate for a submission account. Please contact database staff if you require further information:
e-mail: datasubs@ebi.ac.uk

[index]

Sequence Alignment Submissions

Sequence alignment data (e.g. from phylogenetic and population analysis etc.) of nucleotide sequences can be submitted to the EBI using the WWW submission tool Webin-Align. As an additional service to the scientific community, amino acid alignments are also accepted. Submissions are assigned a number e.g. ALIGN_000001 within two working days of receipt pending review by an EMBL database biologist. We suggest that this number if quoted in any resulting publications.

Alignment data can be retrieved in the following ways:

- EBI FTP serv r: by anonymous FTP from FTP.EBI.AC.UK in directory /pub/databases/embl/align
- EBI File server: by sending an e-mail message to netserv@ebi.ac.uk including the line HELP ALIGN or GET ALIGN:DS8200.DAT;

- EBI WWW pages: ftp://ftp.ebi.ac.uk/pub/databases/embl/align/
- List of alignments is availabl at http://www3.ebi.ac.uk/Services/align/listali.html

Detailed information on how to submit sequence alignment data to the EBI is available.

[index]

Updating your data

Once a database entry has been created from a submission, a copy is sent to the submitter for their reference. Submitters may send comments or corrections using one of the update options described below.
With the passage of time an entry which was correct at the time of creation may become out of date: the authors may make corrections to the sequence itself, or may discover new features which require annotation.
Since such findings are often not published, it is very important that authors communicate their new findings to the database.

Update options:

1. WWW form:

- URL: http://www3.ebi.ac.uk/Services/webin/update/update.html
- This is the preferred option

2. Update form available via anonymous FTP:

- FTP address: pub/databases/embl/release/update.txt
- The completed form should be sent via email to update@ebi.ac.uk

3. Freetext message:

- The message should be sent to update@ebi.ac.uk including the following information: accession number of the sequence to be updated, update information and reason for the update

Citation Updates. Most submissions represent data that have not yet been accepted for publication, and therefore a full journal citation for the data is not available when the entry is created. Adding this information at a later date requires that the database staff identify which submissions correspond to which publications. This task is not always straightforward, for instance, if the accession number is not included in the article, or if the submitted and the published data are not identical. We therefore urge researchers to let us know when and where data they have submitted to us are published, and to include relevant accession numbers in such publications.

[index]

Appendix I. EBI WWW , E-mail and FTP Servers

1. WWW Server

Sequence submission and update via WWW:

http://www.ebi.ac.uk/embl/Submission/webin.html

download Sequin using your WWW browser:

http://www3.ebi.ac.uk/Services/Sequin/

2. **E-mail Server.** Computer users with access to Internet (directly or via a gateway) can obtain copies of the data submission form, or of database entries, by sending commands to a file server at EMBL. The file server facility is provided free of charge, though users may have to meet some or all of the communication costs, depending on the accounting system of their local computer service. To use this facility, send file server commands (as electronic mail) to the address NetServ@EBI.AC.UK . (Please do not use the datalib@ebi.ac.uk. address for this). Each line of the mail message should consist of a single file server command, and nothing else.
The most important file server command, to get users started, is HELP. If the file server receives this command, it will return a help file to the sender, explaining in some detail how to use the facility.

To request help information the mail message should contain the following command:

HELP

To request a copy of the data submission form, your mail message should contain one of the following commands:

GET DOC:DATASUB.TXT

For those requiring software, an extra message, HELP SOFTWARE, will provide relevant information for installation of the programs.
Users can also request specific sequences via the File Server. Information on how to do this is provided in the HELP file.

3.FTP Server

EBI has an anonymous FTP server operational at the Internet address FTP.EBI.AC.UK. (ftp://ftp.ebi.ac.uk/pub/)
Users should log in with the username "anonymous", and for the password give their e-mail address.
The FTP archive currently contains molecular biology databases, free molecular biology software and other files similar to the facilities offered by the e-mail server NetServ@EBI.AC.UK .
Weekly batches of additions to the EMBL nucleotide sequence database (data from EMBL/GenBank/DDBJ) are made available as compressed tar files in the directory: ftp://ftp.ebi.ac.uk/pub/databases/embl/new

[index]

Appendix II. How to contact the Nucleotide Sequence Database

EMBL Nucleotide Sequence Database:

Computer network:

- DATASUBS@EBI.AC.UK (for data submissions);
- DATALIB@EBI.AC.UK (for other enquiries);
- UPDATE@EBI.AC.UK (for updates and notification of publication)

Postal address:

- EMBL Nucleotide Sequence Submissions, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.

Telephone:

- +44-1223-494444 (general)
- +44-1223-494499 (submissions)

Telefax:

- +44-1223-494468 (general)
- +44-1223-494472 (submissions)

• [index]

This page is maintained by support@ebi.ac.uk. Last updated: Monday, December 17, 2001

Submit to GenBank

▸ Submitting Sequence Data to GenBank

The most important source of new data for GenBank® is direct submissions from scientists. GenBank depends on its contributors to help keep the database as comprehensive, current, and accurate as possible. NCBI provides timely and accurate processing and biological review of new entries and updates to existing entries, and is ready to assist authors who have new data to submit.

NOTE: The 'Authorin' submission tool and the E-mail submission form were phased out on December 31, 1998, and submissions made with those tools are no longer accepted as of that date. Instead, please use the improved submission tools, BankIt and Sequin, described below.

▸ *Submit now!!*

Sequin
Stand-alone sequence submission tool

BankIt
For quick and simple submissions

VecScreen
Vector contamination screening tool

▸ *GenBank*

GenBank
overview of the database

Search GenBank
explore the data

▸ Receiving an accession number for your manuscript

Most journals now expect that DNA and amino acid sequences that appear in articles will be submitted to a sequence database before publication. Soon after submission, you will receive an accession number from the database which you will be able to use in your article to refer to the sequence. Please be aware that it is only necessary to submit the sequence to one database, whichever one is most convenient, without regard for where the sequence may be published. Data exchange between GenBank, EMBL and DDBJ occurs daily. Sequence data submitted in advance of publication can be kept confidential if requested.

Below are described various ways of submitting DNA sequences to GenBank. Essentially, there are two principal ways, BankIt and Sequin. BankIt is a Web submission tool and recommended for simple submissions. With BankIt you can indicate coding regions on an mRNA along with a product and gene name. For more control over annotating your entry, segment d r cords, or very long entries, Sequin, a stand-alone

submission tool, is suggested.

GenBank will provide you with an acc ssion number to identify your sequence, usually within two working days, if the submission is received via electronic mail. This accession number serves as confirmation that you have submitted your data, and allows the community to retrieve the data upon reading the journal article.

The accession number should be included in your manuscript, preferably in a footnote on the first page of the article, or as required by individual journal procedures.

▸ BankIt - submitting via the WWW

NCBI has developed a WWW form, called BankIt, for convenient and quick submission of sequence data.

BankIt allows you to enter sequence information into a form, edit as necessary, and add biological annotation (e.g., coding regions, mRNA features). BankIt transforms your data into GenBank format for your review and when your record is completed, it can be submitted directly to GenBank. You have the option of adding information by using text boxes to describe in your own words the source of the sequence and its biological features. The GenBank annotation staff reviews the submitted textual information, incorporates it into the appropriate structured fields, and returns the record by e-mail for your review.

BankIt is compatible with Netscape clients for Unix, Macs, and PCs. In addition, Internet Explorer for the PC and Mac have successfully been used.

▸ Sequin - stand-alone software for the Mac, PC/Windows, and UNIX

If you do not have access to the WWW, NCBI introduces a stand-alone submission program called Sequin.

Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides you through the process of entering your sequence and providing biological and bibliographic annotation. Sequin is designed to simplify the sequence submission process and to provide graphical viewing and editing options. It incorporates robust error checking and accommodates very long sequences and complex annotations.

▸ Special submissions - genomes, batch sequences, alignments

Sequin can be used for the submission of individual or small numbers of sequences. However, it was also designed to facilitate special types of submissions, and should be used

instead of BankIt for the following types of submissions: genomes and oth r very long sequences; multiple sequences such as batch submissions and segmented sets; and population/phylogenetic/mutation studies.

When preparing the submission of a genome, you can import the complete genome sequence into Sequin as well as a file containing the amino acid translations in FASTA format, if available. Sequin will automatically annotate the coding regions intervals based on the translations, and you can use Sequin to make further complex annotations. Sequin can also accept feature annotations in tab-delineated tables. Since the final submission file (*.sqn) will be quite large, please send it to the GenBank staff via FTP rather than by e-mail. To request a temporary FTP directory, please contact genomes@ncbi.nlm.nih.gov.

When preparing a submission that contains multiple sequences, you can import a single file containing all the sequences in FASTA format, or as alignments in FASTA+GAP, PHYLIP, or NEXUS format. In addition, for population/phylogenetic/mutation studies, you can annotate one sequence and propagate the features onto the other sequences. When you complete the submission and select the 'prepare submission' option in the 'File' menu, Sequin will prepare a single *.sqn file that contains all the sequences. Send the *.sqn file by e-mail to:

gb-sub@ncbi.nlm.nih.gov .

If you are submitting two or more Sequin files, each of which contains multiple sequences, send each *.sqn file in a separate e-mail message.

Please refer to the Sequin Quick Guide and documentation for additional information, both of which are accessible from the Sequin Web page.

▸ Sending the Data to GenBank

When using BankIt, the prepared sequence entries are submitted directly to GenBank through the WWW.

When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail to:

gb-sub@ncbi.nlm.nih.gov

As an alternative, the submission file can be copied to floppy disk and mailed to GenBank Submissions at:

GenBank Submissions
National Center for Biotechnology Information
National Library of Medicine

Bldg. 38A, Room 8N-803
Bethesda, MD 20894

Please label the disk with your name and file name and indicate whether it is a PC or MAC disk.

▸ Updates

NCBI processes update requests as well as new submissions. You can provide additional annotation, correct errors or omissions, or request the release of your "hold-until-published" record. BankIt or Sequin may be used for updates, or you can request changes as text in the body of an e-mail message. Be sure to give the accession number of the sequence to be updated along with all update information. Send it to:

update@ncbi.nlm.nih.gov

Submitters of a record maintain editorial control of that record. Any third party update information will be forwarded to the submitters of the record for review. Changes will be made to the record only at the submitters' request. If submitters can no longer be contacted, GenBank reserves the right to edit an entry to agree with the information presented in the original publication (s) cited in the entry.

▸ Submission of ESTs, STSs and GSSs

Batches of ESTs (expressed sequence tags), STSs (sequence tagged sites), and GSSs (genome survey sequences) can be submitted via special streamlined procedures.

▸ Submission of HTGS Records

The NCBI has developed a protocol for high throughput genome sequencing centers to use when they submit large genomic records (usually Cosmids or BACs). Specialized tools, including fa2htgs and a "genome center version" of Sequin, have been created to help such centers produce these submission files in a convenient way. The HTG page not only provides detailed submission instructions to genome centers, but also informs GenBank users how to access the HTG sequences.

▸ Confidentiality

Some authors are concerned that the appearance of their data in GenBank prior to publication will compromise their work. GenBank will, upon request, withhold release of new submissions for a specified period of time. However, if a paper

citing the sequence or accession number is published prior to the specified date, your sequence will be released upon publication.

In order to pr vent the delay in the appearance of published sequence data, we urge authors to inform us of th appearance of the published data. As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address:

update@ncbi.nlm.nih.gov

▸ Submission of SNPs and other polymorphism data

Data on genetic variation in humans and other organisms can be submitted to the NCBI Database of Single Nucleotide Polymorphisms (dbSNP). Entries include single nucleotide polymorphisms (SNPs), small-scale insertion/deletions, polymorphic repetitive elements, and microsatellite variation. dbSNP is a separate resource from the GenBank database, and submissions do not receive GenBank accessions as noted above. However, dbSNP entries do receive dbSNP identifiers and contain links to associated GenBank records. Further information about submitting data is accessible from the sidebar of the dbSNP home page.

Revised January 7, 2002

GenBank Flat File Format

```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      baker's yeast.
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Hemiascomycetes; Saccharomycetales;
            Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrence,C.W.
  TITLE     Cloning and sequence of REV7, a gene whose function is required for
            DNA damage-induced mutagenesis in Saccharomyces cerevisiae
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  MEDLINE   95176709
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requires Axl2p, a novel
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  MEDLINE   96194260
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology, Yale University, New
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
     CDS             <1..206
                     /codon_start=3
                     /product="TCP1-beta"
                     /protein_id="AAA98665.1"
                     /db_xref="GI:1293614"
                     /translation="SSIYNGISTSGLDLNNGTIADMRQLGIVESYKLKRAVVSSASEA
                     AEVLLRVDNIIRARPRTANRQHM"
     gene            687..3158
                     /gene="AXL2"
     CDS             687..3158
                     /gene="AXL2"
                     /note="plasma membrane glycoprotein"
                     /codon_start=1
                     /function="required for axial budding pattern of S.
                     cerevisiae"
                     /product="Axl2p"
                     /protein_id="AAA98666.1"
                     /db_xref="GI:1293615"
                     /translation="MTQLQISLLLTATISLLHLVVATPYEAYPIGKQYPPVARVNESF
                     TFQISNDTYKSSVDKTAQITYNCFDLPSWLSFDSSSRTFSGEPSSDLLSDANTTLYFN
                     VILEGTDSADSTSLNNTYQFVVTNRPSISLSSDFNLLALLKNYGYTNGKNALKLDPNE
```

NCBI Site Map

```
                     VFNVTFDRSMFTNEESIVSYYGRSQLYNAPLPNWLFFDSGELKFTGTAPVINSAIAPE
                     TSYSFVIIATDIEGFSAVEVEFELVIGAHQLTTSIQNSLIINVTDTGNVSYDLPLNYV
                     YLDDDPISSDKLGSINLLDAPDWVALDNATISGSVPDELLGKNSNPANFSVSIYDTYG
                     DVIYFNFEVVSTTDLFAISSLPNINATRGEWFSYYFLPSQFTDYVNTNVSLEFTNSSQ
                     DHDWVKFQSSNLTLAGEVPKNFDKLSLGLKANQGSQSQELYFNIIGMDSKITHSNHSA
                     NATSTRSSHHSTSTSSYTSSTYTAKISSTSAAATSSAPAALPAANKTSSHNKKAVAIA
                     CGVAIPLGVILVALICFLIFWRRRRENPDDENLPHAISGPDLNNPANKPNQENATPLN
                     NPFDDDASSYDDTSIARRLAALNTLKLDNHSATESDISSVDEKRDSLSGMNTYNDQFQ
                     SQSKEELLAKPPVQPPESPFFDPQNRSSSVYMDSEPAVNKSWRYTGNLSPVSDIVRDS
                     YGSQKTVDTEKLFDLEAPEKEKRTSRDVTMSSLDPWNSNISPSPVRKSVTPSPYNVTK
                     HRNRHLQNIQDSQSGKNGITPTTMSTSSSDDFVPVKDGENFCWVHSMEPDRRPSKKRL
                     VDFSNKSNVNVGQVKDIHGRIPEML"
gene                 complement(3300..4037)
                     /gene="REV7"
CDS                  complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
BASE COUNT      1510 a   1074 c    835 g   1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
     1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
     1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
     1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
     1561 acttattgga tgctccagac tgggtggcat tagataatgc taccatttcc gggtctgtcc
     1621 cagatgaatt actcggtaag aactccaatc ctgccaattt ttctgtgtcc atttatgata
     1681 cttatggtga tgtgatttat ttcaacttcg aagttgtctc cacaacggat ttgtttgcca
     1741 ttagttctct tcccaatatt aacgctacaa ggggtgaatg gttctcctac tatttttgc
     1801 cttctcagtt tacagactac gtgaatacaa acgtttcatt agagtttact aattcaagcc
     1861 aagaccatga ctgggtgaaa ttccaatcat ctaatttaac attagctgga gaagtgccca
     1921 agaatttcga caagctttca ttaggtttga aagcgaacca aggttcacaa tctcaagagc
     1981 tatattttaa catcattggc atggattcaa agataactca ctcaaaccac agtgcgaatg
     2041 caacgtccac aagaagttct caccactcca cctcaacaag ttcttacaca tcttctactt
     2101 acactgcaaa aatttcttct acctccgctg ctgctacttc ttctgctcca gcagcgctgc
     2161 cagcagccaa taaaacttca tctcacaata aaaaagcagt agcaattgcg tgcggtgttg
     2221 ctatcccatt aggcgttatc ctagtagctc tcatttgctt cctaatattc tggagacgca
     2281 gaagggaaaa tccagacgat gaaaacttac cgcatgctat tagtggacct gatttgaata
     2341 atcctgcaaa taaaccaaat caagaaaacg ctacaccttt gaacaacccc tttgatgatg
```

NCBI Site Map

```
2401 atgcttcctc gtacgatgat acttcaatag caagaagatt ggctgctttg aacactttga
2461 aattggataa ccactctgcc actgaatctg atatttccag cgtggatgaa aagagagatt
2521 ctctatcagg tatgaataca tacaatgatc agttccaatc ccaaagtaaa gaagaattat
2581 tagcaaaacc cccagtacag cctccagaga gcccgttctt tgacccacag aataggtctt
2641 cttctgtgta tatggatagt gaaccagcag taaataaatc ctggcgatat actggcaacc
2701 tgtcaccagt ctctgatatt gtcagagaca gttacggatc acaaaaaact gttgatacag
2761 aaaaactttt cgatttagaa gcaccagaga aggaaaaacg tacgtcaagg gatgtcacta
2821 tgtcttcact ggacccttgg aacagcaata ttagcccttc tcccgtaaga aaatcagtaa
2881 caccatcacc atataacgta acgaagcatc gtaaccgcca cttacaaaat attcaagact
2941 ctcaaagcgg taaaaacgga atcactccca caacaatgtc aacttcatct tctgacgatt
3001 ttgttccggt taaagatggt gaaaattttt gctgggtcca tagcatggaa ccagacagaa
3061 gaccaagtaa gaaaaggtta gtagattttt caaataagag taatgtcaat gttggtcaag
3121 ttaaggacat tcacggacgc atcccagaaa tgctgtgatt atacgcaacg atattttgct
3181 taattttatt ttcctgtttt atttttttatt agtggtttac agatacccta tattttattt
3241 agtttttata cttagagaca tttaatttta attccattct tcaaatttca tttttgcact
3301 taaaacaaag atccaaaaat gctctcgccc tcttcatatt gagaatacac tccattcaaa
3361 attttgtcgt caccgctgat taatttttca ctaaactgat gaataatcaa aggccccacg
3421 tcagaaccga ctaaagaagt gagttttatt ttaggaggtt gaaaaccatt attgtctggt
3481 aaattttcat cttcttgaca tttaacccag tttgaatccc tttcaatttc tgcttttcc
3541 tccaaactat cgaccctcct gtttctgtcc aacttatgtc ctagttccaa ttcgatcgca
3601 ttaataactg cttcaaatgt tattgtgtca tcgttgactt taggtaattt ctccaaatgc
3661 ataatcaaac tatttaagga agatcggaat tcgtcgaaca cttcagtttc cgtaatgatc
3721 tgatcgtctt tatccacatg ttgtaattca ctaaaatcta aaacgtattt ttcaatgcat
3781 aaatcgttct ttttattaat aatgcagatg gaaaatctgt aaacgtgcgt taatttagaa
3841 agaacatcca gtataagttc ttctatatag tcaattaaag caggatgcct attaatggga
3901 acgaactgcg gcaagttgaa tgactggtaa gtagtgtagt cgaatgactg aggtgggtat
3961 acatttctat aaaataaaat caaattaatg tagcatttta agtatacccc cagccacttc
4021 tctacccatc tattcataaa gctgacgcaa cgattactat tttttttttc ttcttggatc
4081 tcagtcgtcg caaaaacgta taccttcttt ttccgacctt ttttttagct ttctggaaaa
4141 gtttatatta gttaaacagg gtctagtctt agtgtgaaag ctagtggttt cgattgactg
4201 atattaagaa agtggaaatt aaattagtag tgtagacgta tatgcatatg tatttctcgc
4261 ctgtttatgt ttctacgtac ttttgattta tagcaagggg aaaagaaata catactattt
4321 tttggtaaag gtgaaagcat aatgtaaaag ctagaataaa atggacgaaa taaagagagg
4381 cttagttcat ctttttttca aaaagcaccc aatgataata actaaaatga aaaggatttg
4441 ccatctgtca gcaacatcag ttgtgtgagc aataataaaa tcatcacctc cgttgccttt
4501 agcgcgtttg tcgtttgtat cttccgtaat tttagtctta tcaatgggaa tcataaattt
4561 tccaatgaat tagcaatttc gtccaattct ttttgagctt cttcatattt gctttggaat
4621 tcttcgcact tcttttccca ttcatctctt tcttcttcca aagcaacgat ccttctaccc
4681 atttgctcag agttcaaatc ggcctctttc agtttatcca ttgcttcctt cagtttggct
4741 tcactgtctt ctagctgttg ttctagatcc tggtttttct tggtgtagtt ctcattatta
4801 gatctcaagt tattggagtc ttcagccaat tgctttgtat cagacaattg actctctaac
4861 ttctccactt cactgtcgag ttgctcgttt ttagcggaca aagatttaat ctcgttttct
4921 ttttcagtgt tagattgctc taattctttg agctgttctc tcagctcctc atatttttct
4981 tgccatgact cagattctaa ttttaagcta ttcaatttct ctttgatc
//
```

Other Formats: FASTA | ASN.1

Back to Top

Examples of other records that show a range of biological features

---

FIELD COMMENTS

NCBI Site Map

LOCUS

- **Locus Name**

The locus name was originally designed to help group entries with similar sequences: the first three characters usually designated the organism; the fourth and fifth characters were used to show other group designations, such as gene product; for segmented entries the last character was one of a series of sequential integers. (See GenBank release notes section 3.4.4 for more info.)

However, the ten characters in the locus name are no longer sufficient to represent the amount of information originally intended to be contained in the Locus name. The only rule now applied in assigning a Locus name is that it must be unique. For example, for GenBank records that have 6-character accessions (e.g., U12345), the locus name is usually the first letter of the genus and species name followed by the accession number. For 8-character character accessions (e.g., AF123456), the locus name is just the accession number.

The RefSeq database of reference sequences assigns formal locus names to each record, based on gene symbol. RefSeq is separate from the GenBank database, but contains cross-references to corresponding GenBank records.

Entrez Search Field: Accession Number [ACCN]
Search Tip: It is better to search for the actual accession number rather than the locus name, since the accessions are stable and locus names can change.

- **Sequence Length**

Number of nucleotide base pairs (or amino acid residues) in the sequence record.

There is no maximum limit on the size of a sequence that can be submitted to GenBank - you can submit a whole genome if you have a contiguous piece of sequence from a single molecule type. However, there is a limit of 350 kb on an individual GenBank record (with some exceptions, as noted in section 1.3.2 of the release notes for GenBank 112.0). That limit was agreed upon by the international collaboratoring

sequence databases to facilitate handling of sequence data by various software programs. (For more information, see NCBI News articles on Complete Genomes and GenBank Enters Megabase Era.) The minimum length required for submission is 50 bp, although there might be some shorter records from past years.

Entrez Search Field: Sequence Length [SLEN]
Search Tips: (1) The current version of Entrez requires that seqeunce length be written as six digits, e.g., 150 bp = 000150. The upcoming release of Entrez will not require that. (2) To retrieve records within a range of lengths, use the colon as the range operator, e.g., 002500:002600[slen]. (3) To retrieve all sequences shorter than a certain number, use 000002 as the lower bound, e.g., 000002:000100 [slen]. (4) To retrieve all sequences longer than a certain number, use 999999 as the upper bound, e.g., 325000:999999[slen].

- Molecule Type

The type of molecule that was sequenced.

Each GenBank record must contain contiguous sequence data from a single molecule type. The various molecule types are described in the Sequin documentation, and can include genomic DNA, genomic RNA, precursor RNA, mRNA (cDNA), ribosomal RNA, transfer RNA, small nuclear RNA, and small cytoplasmic RNA.

Entrez Search Field: Properties [PROP]
Search Tip: Search term should be in the format: **biomol_genomic**, **biomol_mRNA**, etc. For more examples, view search the Properties field in "List Terms" mode to view the index.

- GenBank Division

The GenBank database is divided into 16 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences

11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTGS sequences (high throughput genomic sequences)

For more information, see section 3.3 of the GenBank release notes.

The RNA division of GenBank was removed in release 113.0 (August 1999). Sequences that were previously in the RNA division have been moved to the appropriate organismal division. (See section 1.3.2 of the GenBank 113.0 release notes for additional information.)

The CON division was added in release 115.0 (December 1999). Records in that division contain no sequence data. Instead, they contain instructions on how to construct contigs from multiple GenBank records. See the Fall 1999 NCBI News and section 1.3.3 of GenBank 115.0 release notes for details. The CON division is not listed above because it is still experimental.

Entrez Search Field: Properties [PROP]
Search Tip: Search term should be in the format: gbdiv_pri, gbdiv_est, etc. For more examples, view search the Properties field in "List Terms" mode to view the index. To eliminate all sequences from a particular division, you can use a Boolean query such as: human[orgn] NOT gbdiv_est[prop]

- Modification Date

The date in the LOCUS field is the **date of last modification.** In some cases, it might correspond to the release date, but there is no way to tell just by looking at the record. If you need to know the first date of public availability for a specific sequence record, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you, and let you know the date of first public release. If the sequence was originally submitted to our collaborators at DDBJ or EMBL, rather than to GenBank, we will ask them to send the release date information to you. (See also notes re: date in the Direct Submission reference.)

Entrez Search Field: Modification Date [MDAT]

Search Tips: (1) Enter search term in the format: yyyy/mm/dd, e.g., 1999/07/25. (2) To retrieve records modified between two dates, use the colon as a range operator, e.g., 1999/07/25:1999/07/31[mdat]. (3) You can use the Publication Date [PDAT] field of Entrez to limit search results by the date on which records were added to the Entrez system. Publication date can be ranged just like the Modification Date.

DEFINITION

Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region (CDS), description may be followed by a completeness qualifier, such as "complete cds." (See GenBank release notes section 3.4.5 for more info.)

Entrez Search Field: Title Word [TITL]
Search Tip: Although nucleotide definition lines follow a structured format, GenBank does not use a controlled vocabulary and authors determine the content of their records. Therefore, if a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such synonyms, full spellings, or abbreviations. The 'related records' (or 'neighbors') function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

ACCESSION

The unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345), or two letters followed by six digits (e.g., AF123456). Accession numbers do not change, even if information in the record is changed at the author's request. Sometimes, however, an original accession number might become secondary to a newer accession number, if the authors make a new submission that combines previous sequences, or if for some reason a new submission supercedes an earlier record.

Records from the RefSeq database of reference sequences have a different accession number format that begins with two letters followed by

an underscore bar and six digits:

NT_123456	constructed genomic contigs
NM_123456	mRNAs
NP_123456	proteins
NC_123456	chromosomes

Note: compare accession number with Sequence Identifiers such as Version and GI for nucleotide sequences, and ProteinID and GI for amino acid sequences.

Entrez Search Field: Accession [ACCN]
Search Tip: The letters in the accession number can be written in upper or lower case. RefSeq accessions must contain an underscore bar between the letters and the numbers, e.g, NM_002111.

VERSION

A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. This identification number uses the accession.version format implemented by GenBank/EMBL/DDBJ in February 1999.

If there is any change to the sequence data (even a single base) the version number will be increased, e.g., U12345.1 --> U12345.2, but the accession portion will remain stable.

The accession.version system of sequence identifiers runs parallel to the GI number system. That is, when any change is made to a sequence, it receives a new GI number AND an increase to its version number.

For more information, see section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

A Sequence Revision History tool is available to track the various gi numbers, version numbers, and update dates for sequences that appeared in a specific GenBank record (more information and example).

Entrez Search Field: Can use either Accession [ACCN] or UID

- GI

"GenInfo Identifier" sequence identification

number, in this case, for the nucleotide sequence. If a sequence changes in any way, a new GI number will be assigned.

A separate GI number is also assigned to each protein translation within a nucleotide sequence record, and a new GI is assigned if the protein translation changes in any way (see below).

GI sequence identifiers run parallel to the new **accession.version** system of sequence identifiers. For more information, see the description of Version, above, and section 3.4.7 of the current GenBank release notes.

Entrez Search Field: UID

KEYWORDS

Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period.

The Keyword field is present in sequence records primarily for historical reasons, and is not based on a controlled vocabulary. Keywords are generally present in older records. They are **not** included in newer records unless (1) they are not redundant with any feature, qualifier, or other information present in the record, or (2) the submitter specifically asks for them to be added, and (1) is true, or (3) the sequence needs to be tagged as an EST, STS, GSS or HTG.

Entrez Search Field: Keyword [KYWD]
Search Tip: Since keywords are not present in many records, it is best not to search that field. Instead, search All Fields [ALL], the Text Word [WORD] field, or the Title Word [TITL] field, for progressively narrower retrieval.

SOURCE

Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type. (See section 3.4.10 of the GenBank release notes for more info.)

Entrez Search Field: Organism [ORGN]
Search Tip: For some organisms that have well established common names, such as baker's yeast, mouse, and human, a search for the common name will yield the same results as a search for the scientific name. E.g., a search for "baker's

yeast" in the organism field retrieves the same number of documents as "Saccharomyces cerevisiae." This is true because the Organism field is connected to the NCBI Taxonomy Database, which contains cross-references between common names, scientific names, and synonyms for organisms represented in the Sequence databases.

- Organism

The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database. If the complete lineage of an organism is very long, an abbreviated lineage will be shown in the GenBank record and the complete lineage will be available in the Taxonomy Database. (See also the /db_xref=taxon:nnnn Feature qualifer, below.)

Entrez Search Field: Organism [ORGN]
Search Tip: You can search the Organism field by any node in the taxonomic hierarchy. E.g., you can search for the term "Saccharomyces cerevisiae," "Saccharomycetales," "Ascomycota," etc. to retrieve all the sequences from organisms in a particular taxon.

REFERENCE

Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first.

Some sequences have not been reported in papers and show a status of "unpublished" or "in press." When an accession number and/or sequence data has appeared in print, sequence authors should send the complete citation of the article to update@ncbi.nlm.nih.gov and the GenBank staff will revise the record.

Various classes of publication can be present in the References field, including journal article, book chapter, book, thesis/monograph, proceedings chapter, proceedings from a meeting, and patent. The last citation in the References field contains information about the submission itself, rather than a literature citation (see Direct Submission, below).

Entrez Search Field: The various subfields under References are searchable in the Entrez search fields noted below.

- Authors

List of authors in the order in which they appear in the cited article.

Entrez Search Field: Author [AUTH]
Search Tip: Enter author names in the form: Lastname AB (without periods after the initials). Initials can be omitted. Truncation can also be used to retrieve all names that begin with a character string, e.g., Richards* or Boguski M*.

- Title

Title of the published work, or tentative title of an unpublished work.

Entrez Search Field: Text Word [WORD]
Note: For sequence records, the Title Word [TITL] field of Entrez searches the Definition Line, not the titles of references listed in the record. Therefore, use the Text Word field to search the titles of references (and other text-containing fields).
Search Tip: If a search for a specific term does not retrieve the desired records, try other terms that authors might have used, such synonyms, full spellings, or abbreviations. The 'related records' (or 'neighbors') function of Entrez also allows you to broaden your search by retrieving records with similar sequences, regardless of the descriptive terms used by the submitters.

- Journal

MEDLINE abbreviation of the journal name. (Full spellings can be obtained from the PubMed Journal Browser.)

Entrez Search Field: Journal Name [JOUR]
Search Tip: Journal names can be entered as either the full spelling or the MEDLINE abbreviation. You can search the Journal Name field in "List Terms" mode to view the index of that field, and to select one or more journal names for inclusion in your search.

- MEDLINE

MEDLINE unique identifier (UID).

References that include MEDLINE UIDs contain links from the sequence record to the corresponding MEDLINE record. Conversely, MEDLINE records that contain accession number(s) in the SI (secondary source identifier) field contain links back to the sequence record(s).

Entrez Search Field: It is not possible to search the Nucleotide or Protein sequence databases by MEDLINE UID. However, you can search the Literature (PubMed) database of Entrez for the MEDLINE UID, and then link to the associated sequence records.

- Direct Submission

Contact information of the submitter, such as institute/department and postal address. This is always the The last citation in the References field. Some older records do not contain the "Direct Submission" reference. However, it is required in all new records.

The Authors subfield contains the submitter name (s), Title contains the words "Direct Submission," and Journal contains the address.

The date in the Journal subfield is the date on which the author prepared the submission. In many cases, it is also the date on which the sequence was received by the GenBank staff, but it is not the date of first public release. If you need to know the latter, send a message to info@ncbi.nlm.nih.gov. We will check the history of the record for you.

Entrez Search Field: Use the Author Field [AUTH] if searching for the author name. Use All Fields [ALL] if searching for an element of the author's address (e.g., Yale University). Note, however, that retrieved records might contain the institution name in a field such as Comment, rather than in the Direct Submission reference, so you might get some false hits.
Search Tip: It is sometimes helpful to search for both the full spelling and an abbreviation, e.g., "Washington University" OR "WashU", since the spelling used by authors might vary.

FEATURES

Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features. (See section 3.4.12 of the GenBank release notes for more info.)

A **complete list of features** is available in three places:

- Appendix III: Feature keys reference of the DDBJ/EMBL/GenBank Feature Table provides definitions, optional qualifiers, and comments for each feature. An alphabetical list is also available. Appendix IV: Summary of qualifiers for feature keys provides definitions for the Feature qualifiers.
- Sequin Help documentation (scroll down to 'Features' in the table of contents to see an alphabetical list of features with links to descriptions)
- section 3.4.12.1 of the GenBank release notes

The **location of each feature** is provided as well, an can be a single base, a contiguous span of bases, a joining of sequence spans, and other representations. If a feature is located on the complementary strand, the word "complement" will appear before the base span. If the "<" symbol precedes a base span, the sequence is partial on the 5' end (e.g., CDS <1..206). If the ">" symbol follows a base span, the sequence is partial on the 3' end (e.g., CDS 435..915>).

For more information about feature locations, see the Sequin Help Documentation and section 3.4.12.2 of the GenBank release notes.

Entrez Search Field: Feature Key [FKEY]
Search Tip: To scroll through the list of available features, search the Feature Key field in List Terms mode. You can then select one or more features from the list to include in your query. For example, you can limit your search to records that contain both primer_bind and promoter features.

- Source

Mandatory feature in each record that summarizes the length of the sequence, sci ntific name of

the source organism, and Taxon ID number. Can also include other information such as map location, strain, clone,tissue type, etc., if provided by submitter.

Entrez Search Field: All Fields [ALL] can be used to search for some elements in the source field, such as strain, clone, tissue type.

Use the Sequence Length [SLEN] field to search by length, and the Organism [ORGN] field to search by organism name.

Since map location is written as free text and can be represented in a number of ways (e.g., chromosome number, cytogenetic location, marker name, physical map location), it is not directly searchable in the Entrez nucleotides or proteins databases. However, there are a number of resources that allow you to browse and/or search the maps of various genomes.

Taxon

A stable unique identification number for the taxon of the source oganism. A taxonomy ID number is assigned to each taxon (species, genus, family, etc.) in the NCBI Taxonomy Database. See also the Organism field, above.

Entrez Search Field: The Taxonomy ID number is not seachable in the Organism search field of Entrez, but is searchable in the Taxonomy Browser

Note: The /db_xref qualifier is one of many that can be applied to various features. A complete list is available in Appendix IV: Summary of qualifiers for feature keys of the DDBJ/EMBL/GenBank Feature Table, and in section 3.4.12.3 of the GenBank release notes. Appendix III: Feature keys reference shows which qualifiers can be used with specific features (see alphabetical list).

- **CDS**

Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons). The CDS feature includes an amino acid translation. Authors can specify the nature of the CDS by using the qualifier /evidence=experimental

or /evidence=not_experimental.

Submitters are also encouraged to annotate the mRNA feature, which includes the 5'untranslated region (5'UTR), coding sequences (CDS, exon) and 3'untranslated region (3'UTR).

Entrez Search Field: Feature Key [FKEY]
Search Tip: You can use this field to limit your search to records that contain a particular feature, such as CDS. To scroll through the list of available features, search the Feature Key field in List Terms mode. A complete list of features is also available from the resources noted above.

Protein ID

A protein sequence identification number in the accession.version format that was implemented by GenBank/EMBL/DDBJ in February 1999 (see Version for additional information). Protein IDs consist of three letters followed by five digits, a dot, and a version number. If there is any change to the sequence data (even a single amino acid), the version number will be increased, but the accession portion will remain stable (e.g., AAA98665.1 will change to AAA98665.2).

Entrez Search Field: Can use either the Accession [ACCN] or UID field of the Entrez Proteins database.

GI

"GenInfo Identifier" sequence identification number, in this case, for the protein translation.

The GI system of sequence identifiers runs parallel to the **accession.version** system, which was implemented by GenBank, EMBL, and DDBJ in February 1999. Therefore, if the protein sequence changes in any way, it will receive a new GI number, and the suffix of the Protein ID will be incremented by one.

For more information, see the description of Protein ID, above, section 1.3.2 of the GenBank 111.0 release notes, and section 3.4.7 of the current GenBank release notes.

Entrez Search Field: Use the UID field of the **Entrez Proteins** database (the UID field of the

Entrez Nucleotides database should be used only for nucleotide sequence identifiers).

Translation

The amino acid translation corresponding to the nucleotide coding sequence (CDS). In many cases, the translations are conceptual. Note that authors can indicate whether the CDS is based on experimental or non-experimental evidence.

Entrez Search Field:It is not possible to search the translation subfield using Entrez. If you want use a string of amino acids as a query to retrieve similar protein sequences, use BLAST instead.

- **Gene**

A region of biological interest identified as a gene and for which a name has been assigned. The base span for the gene feature is dependent on the furthest 5' and 3' features. Additional examples of records that show the relationship between gene features and other features such as mRNA and CDS are AF165912 and AF090832.

Entrez Search Field: Feature Key [FKEY]
Search Tip: You can use this field to limit your search to records that contain a particular feature, such as gene. To scroll through the list of available features, search the Feature Key field in List Terms mode. A complete list of features is also available from the resources noted above.

---

complement

Indicates the feature is located on the complementary strand.

- **Other Features**

Examples of other records that show a variety of biological features; a graphic format is also available for each sequence record, and visually represents the annotated features:

- **AF165912** (gene, promoter, TATA signal, mRNA, 5'UTR, CDS, 3'UTR) GenBank flat file
- **AF090832** (protein bind, gene, 5'UTR, mRNA,

CDS, 3'UTR) GenBank flat file
- **L00727** (alternatively spliced mRNAs) GenBank flat file)

A complete list of features is available from the resources noted above.

BASE COUNT

The number of A, C, G, and T bases in a sequence.

ORIGIN

The ORIGIN may be left blank, may appear as 'Unreported,' or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). This information is only present in older records.

The sequence data begin on the line immediately below Origin. To view/save the sequence data only, display the record in FASTA format. More information about the FASTA format is accessible from the BLAST Web pages.
